学校编码：10384　　分类号_____密级_____
学　　号：23020121152930　　UDC_____

硕　士　学　位　论　文

基于降维的基因表达数据分类算法研究

Classification of Gene Expression Data Based on Dimension Reduction

李　亮　亮

指导教师姓名：　闵　小　平　副教授
专　业　名　称：　计算机系统结构
论文提交日期：　2015　年　4　月
论文答辩时间：　2015　年　　月
学位授予日期：　2015　年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2015　年　　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

利用基因芯片技术能够做到同时对多到数以万计的基因进行并行分析，随着该技术越来越成熟并得到广泛应用，现在已经有越来越多的基因表达数据测定出来，亟需处理，借助于计算机工具以及机器学习方法对这些大量数据进行分析是现在一个很重要的研究领域。其中，基因表达数据的分类研究是该领域的一个热点，通过分类能够帮助研究者发现正常细胞组织与疾病组织之间基因的本质差异，识别致病基因，对基因型疾病的临床诊断和治疗具有重要的意义。

基因表达数据具有“样本少、维数高、分布不平衡”的特点，这给分类带来了很多的困难和挑战。目前解决此问题的一个有效方法是在分类前对高维数据进行特征提取和降维，以达到去除与分类无关的基因、降低计算复杂度、提高分类准确率的目的。

本文首先分别用 PCA、ReliefF、LLE 和 Isomap 几种降维算法对原始数据进行降维，然后对降维后的基因表达数据用朴素贝叶斯分类算法进行分类，并通过分类结果比较了不同降维方法的性能。然后在此基础上，本文提出了 RLLE（relevant component based LLE）降维算法，把 ReliefF 特征提取与 LLE 降维相结合，试验结果表明，RLLE 算法的降维效果要好于传统的几种降维方法。

Alex Rodriguez, Alessandro Laio 提出的“基于快速寻找密度峰值的聚类算法”是一种很简洁且聚类效果很好的聚类算法，但是该算法对“样本少，维数高”的基因表达数据的聚类效果并不是很好，不能分离出正确数目的类中心。本文采用 mRMR 特征提取算法提取原始数据中排名靠前的特征达到降维目的，然后对降维后的基因表达数据重新进行聚类，能够较准确地分离出正确数目的类中心。在此基础上，将该聚类算法发展为有监督特征提取的分类算法：用训练集进行特征提取，再将训练集和测试集合并并聚类，最后根据聚类结果判定测试样本的类别。分类结果表明，基于 mRMR-快速聚类分类算法的分类准确率要好于 k 近邻分类和朴素贝叶斯分类。

关键字：基因表达数据；降维；分类

Abstract

The invention of Gene Chip technique makes it possible to analyse thousands of genes simultaneously. With the rapid development of the technique, more and more gene expression data has been gained and accumulated. It is an important research field to analyse these data using computers and method of machine study, among which classification of gene expression data is a hot field. Classification can help researchers find the essential difference between healthy cell tissues and pathological cell tissues, and then identify malgenic genes, which is of great importance in clinical diagnosis and treatment of disease.

Gene expression data has some features, such as small-sample size，high dimensions and imbalanced distribution, which brings many difficulties and challenges to classification. One effective method is feature-selection and dimension reduction in order to get rid of irrelevant genes, reducing calculating time and then improve accuracy of classification.

This paper reduces the dimension of gene expression data respectively using different algorithms, such as PCA, ReliefF, LLE and Isomap, then classify the data based on Naive Bayes and compare the performance of different algorithms by the classification results. Then we propose RLLE(relevant component based LLE) reduction dimension algorithm by combining ReliefF and LLE. The result of experiment shows LLE is better than traditional reduction dimension algorithm in the aspect of performance.

"Clustering by fast search and find of density peaks" published in the journal of Science by Alex Rodriguez and Alessandro Laio is a clustering algorithm with concise form and good clustering result. However, due to the special features of the data, the algorithm is not able to separate cluster centers of right numbers. We carry out mRMR feature selection algorithm on the data before clustering, then the algorithm is able to find enough centers as well as gaining better classification result. Then we propose a classification algorithm based on the clustering algorithm: first select features from

training data set using mRMR, then merge training and testing data seta of selected features, at last cluster and obtain category of testing samples according to the clustering result. The classification result shows that the algorithm has better performance than k-NN and Naive Bayes classification.

Key Words: Gene Expression Data; Dimension Reduction; Classification

目录

Contents

第一章 绪论

1.1 研究背景

1990 年 10 月 1 日，美国启动跨世纪的“人类基因组计划（Human Genome Project, HGP）”，目的在于全部测定出人类基因组的 30 亿个碱基对序列，找出人类所有基因及其在染色体的位置，最终破解人类的遗传信息。2003 年，通过包括我国在内的六个国家科学家的努力，该计划提前顺利完成，自此人类进入了后基因组时代。后基因组时代的中心研究为基因功能的研究[1,2]，即根据已测定的基因序列来发现基因的生物学功能，将基因序列转化为有用的信息，服务全人类。然而，由于基因序列数据的数量非常庞大，只依靠传统的生物学实验已经无法获取和解释隐藏在这些序列中的有用信息，因此如何去获取这些数量巨大的基因的生物学功能也就成了全世界生物学家共同的难题。为此，发展一种能对大量遗传信息同时进行高效、准确分析的新型测序方法对后基因组时代的基因功能分析是一个迫在眉睫的要求。基因芯片技术就是在这样的科学要求背景下产生的，它的出现为解决此类问题提供了光辉的前景。

1.1.1 基因芯片技术概述

基因芯片[3](gene chip)又称为 DNA 微阵列、DNA 芯片，是根据核酸杂交原理由大量寡核苷酸或 cDNA 探针在固相载体上按照某种特定排列方式密集排列形成的探针阵列。基因芯片技术技术诞生于 20 世纪 80 年代末，是结合利用了分子生物学、微机械学、统计学和计算机科学等多种学科发展起来的一种先进技术，实现了连续化、集成化地处理和分析生物科学样品。

基因芯片技术主要包括以下四个步骤[4]：

1、样品制备

基因芯片中的样品指对照组和实验组中的 mRNA 或总 RNA 样品。样本制备是将 mRNA 或总 RNA 样品分别进行逆转录生成 cDNA，然后将对照组和实验组 cDNA 分别标记 Cy3 和 Cy5 荧光信号。

2、芯片的设计与制备

基因芯片制备是将寡核苷酸片段或 cDNA 作为探针按某种特定序列密集排列在载体上，载体一般选择硅片或玻璃瓶。由于探针很小，为了准确且快速地把探针放到指定的位置，芯片制备还会用到机器人技术。目前的制备方法主要采用点样法或原位合成法。

3、芯片杂交

用 Cy3 和 Cy5 荧光标记的对照组和实验组的 cDNA 等量混合，与芯片上的探针进行杂交。这个杂交过程与一般的分子杂交过程基本相同。为减少错配率，还要对杂交反应的条件进行优化，优化主要与 GC 碱基含量、探针长度和芯片的类型有关。

4、信号检测和结果分析

目前主要采用激光扫描仪，分别用 532nm 和 635nm 波长激光扫描芯片，对于每张芯片，得到 Cy3 和 Cy5 通道两幅图像。采用专门软件，对图像进行分析获取每个点的数字信号，得到原始数据表。另外，还需要对数据进行校正和筛选：对 Cy5 和 Cy3 信号进行校正，消除试验或扫描等各环节因素对数据的影响，同时利用筛选规则对数据中的“坏点”，“小点”和“低信号点”进行筛选，并作标记。由于基因芯片获取的数据较多，信息量大，对于杂交数据的分析、处理、查询和比较等需要一个标准的数据格式。

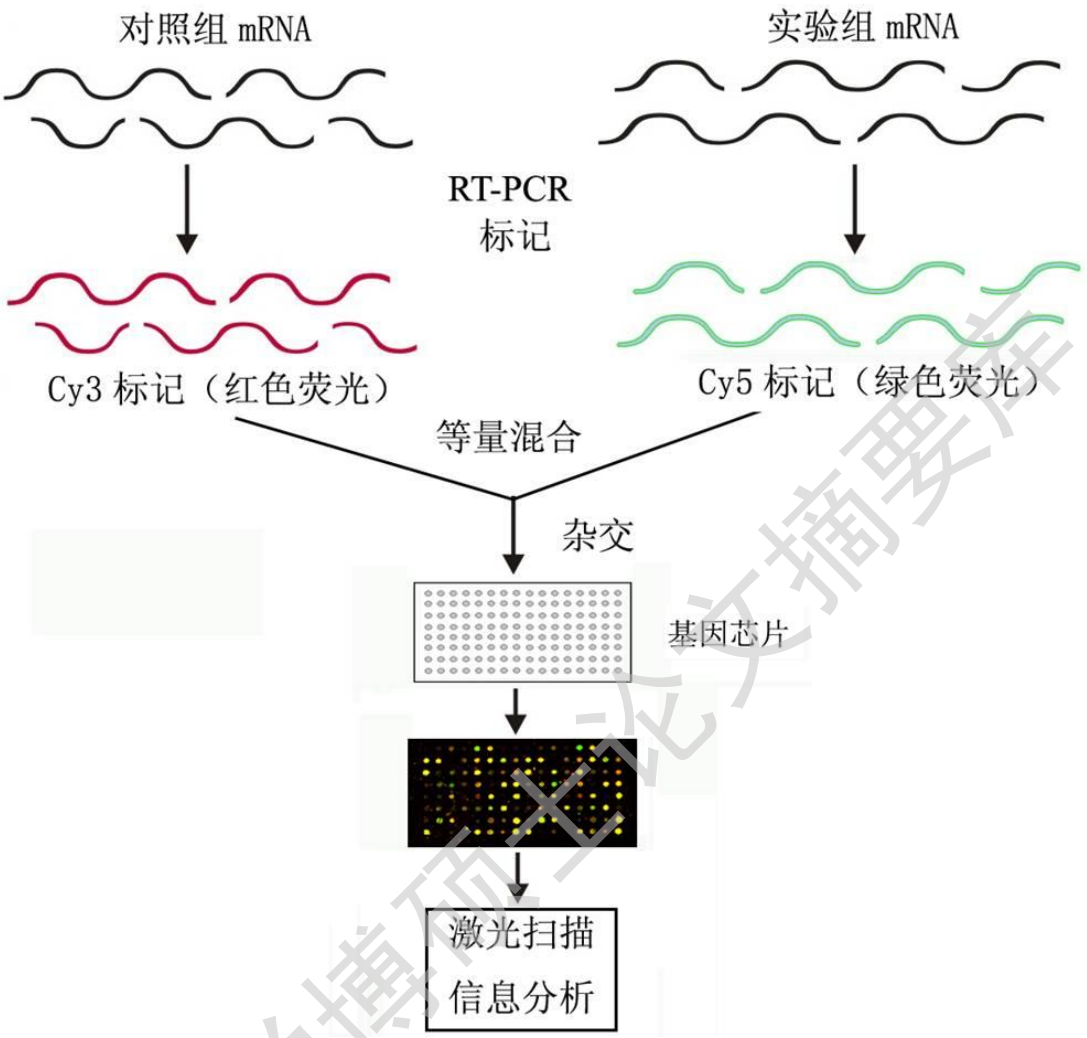

图 1-1 基因芯片技术流程

基因芯片在很多领域都得到了广泛的应用。农业上，能够用于筛选基因突变的农作物，寻找高产、抗病、抗虫、经济价值高的作物以及进行农药的筛选等；环境保护上，能够寻找某些保护基因，用于开发具有防治危害的基因工程药品；司法上，能够进行血型、亲子鉴定以及 DNA 指纹图谱分析等。

基因芯片技术还为解决肿瘤分类问题提供了一种与传统方法不同的新思路[5, 6]。近些年来，利用基因表达数据来筛选分子标记已经成为了比较常用的方法。这方面的研究常常需要用到数学中的统计或结算方法来选择目的基因，对疾病进行分类已经成为了疾病诊断的重要手段。

1.1.2 基因表达数据

在基因芯片上测量不同样本数以万计的基因在不同基因条件下所表达的一组数据即为基因表达数据，这些数据以矩阵的形式存储，因此又称为基因表达矩阵或微阵列数据。矩阵中的行表示样本，列表示基因，某一行的数据代表的是某个样本中不同基因的表达水平，而某一列的数据代表的是某个基因在不同样本中的表达水平。如下所示：

$$X_{n\times p}=\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \quad (1\text{-}1)$$

式中，数据点的总数量为 n*p 个，其中，n 表示样本的数量，p 表示基因的数量，矩阵中的元素值 x_{ij} 则表示第 i 个样本中第 j 个基因的表达值。近些年来，得益于基因表达矩阵的飞速发展，使得同时测定数以万计个基因的表达量成为可能，为生物科研工作者进行基因功能的研究提供了强有力的工具。通过对基因表达数据进行数学方法的研究和比较，并在此基础上进行推测，可以找到蕴藏在基因背后所要表达的重要生物学信息。

基因表达数据通常具有以下特点：

1、样本数和基因数不平衡

一般情况下样本有几十个或几百个，而基因却有成千上万个，体现在基因表达矩阵中就是矩阵的维数非常高，将会给分析带来巨大困难，甚至导致“维数灾难”。

2、数据中包含有大量噪音

这些噪音包括生物学上的噪音和技术上的噪音。

3、数据中包含大量和分类不相关的基因

这些基因的存在不仅干扰了分类的精度而且增加了计算量。